

Using Machine Learning to Estimate Unobserved COVID-19 Infections in North America

Shashank Vaid, PhD, Caglar Cakan, PhD(Candidate), and Mohit Bhandari, MD, PhD

Background: The detection of coronavirus disease 2019 (COVID-19) cases remains a huge challenge. As of April 22, 2020, the COVID-19 pandemic continues to take its toll, with >2.6 million confirmed infections and >183,000 deaths. Dire projections are surfacing almost every day, and policymakers worldwide are using projections for critical decisions. Given this background, we modeled unobserved infections to examine the extent to which we might be grossly underestimating COVID-19 infections in North America.

Methods: We developed a machine-learning model to uncover hidden patterns based on reported cases and to predict potential infections. First, our model relied on dimensionality reduction to identify parameters that were key to uncovering hidden patterns. Next, our predictive analysis used an unbiased hierarchical Bayesian estimator approach to infer past infections from current fatalities.

Results: Our analysis indicates that, when we assumed a 13-day lag time from infection to death, the United States, as of April 22, 2020, likely had at least 1.3 million undetected infections. With a longer lag time—for example, 23 days—there could have been at least 1.7 million undetected infections. Given these assumptions, the number of undetected infections in Canada could have ranged from 60,000 to 80,000. Duarte's elegant unbiased estimator approach suggested that, as of April 22, 2020, the United States had up to >1.6 million undetected infections and Canada had at least 60,000 to 86,000 undetected infections. However, the Johns Hopkins University Center for Systems Science and Engineering data feed on April 22, 2020, reported only 840,476 and 41,650 confirmed cases for the United States and Canada, respectively.

Conclusions: We have identified 2 key findings: (1) as of April 22, 2020, the United States may have had 1.5 to 2.029 times the number of reported infections and Canada may have had 1.44 to 2.06 times the number of reported infections and (2) even if we assume that the fatality and growth rates in the unobservable population (undetected infections) are similar to those in the observable population (confirmed infections), the number of undetected infections may be within ranges similar to those described above. In summary, 2 different approaches indicated similar ranges of undetected infections in North America.

Level of Evidence: Prognostic Level V. See Instructions for Authors for a complete description of levels of evidence.

The detection of coronavirus disease 2019 (COVID-19) cases remains a huge challenge¹. As of April 22, 2020, the COVID-19 pandemic continues to take its toll, with close to 2.6 million confirmed infections and >183,000 deaths². Dire projections are surfacing almost every day, and policymakers worldwide are using projections for critical decisions. While social distancing now appears to be globally accepted, approaches vary substantially. Whereas Hong Kong and Singapore are experimenting with “suppress and lift” measures³, India has been estimated to be at the top of the lockdown stringency index⁴. Intelligence on the number of infections and projected courses has never been more urgent as the world

economy heads toward a contraction of 3% in 2020 and the world faces the worst recession since the Great Depression¹.

While organizations such as the World Health Organization (WHO) are establishing COVID-19-detection protocols⁵, leading scientific opinion and commentaries appear to be highlighting the possibility of detection bias⁶. There also appears to be a grudging acceptance that identifying and quantifying such bias may depend largely on the number of reported cases. The challenge with reported cases is that they are dependent on the extent of testing. As of April 22 2020, the numbers of tests per 1 million population varied greatly across some of the key jurisdictions most impacted by the pandemic,

Disclosure: The authors indicated that no external funding was received for any aspect of this work. On the **Disclosure of Potential Conflicts of Interest** forms, which are provided with the online version of the article, one or more of the authors checked “yes” to indicate that the author had a relevant financial relationship in the biomedical arena outside the submitted work (<http://links.lww.com/JBJS/F886>).

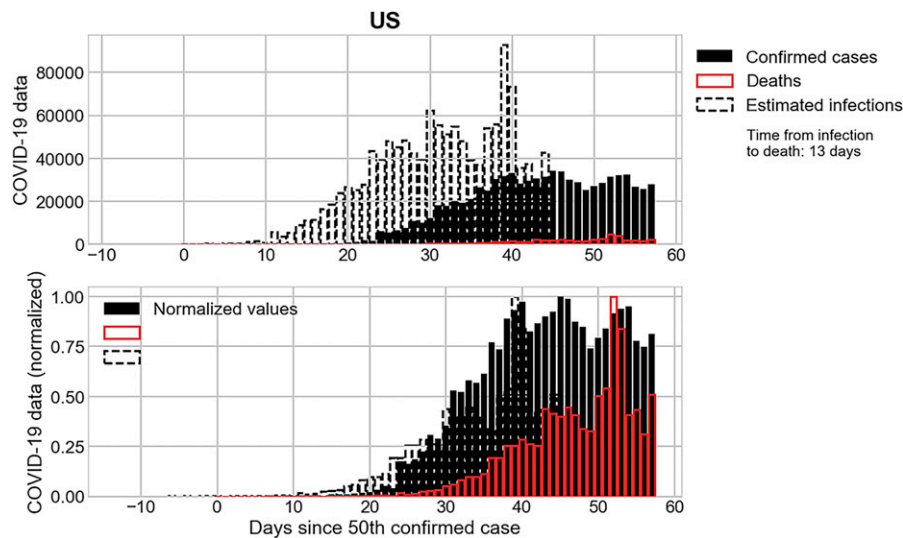


Fig. 1-A

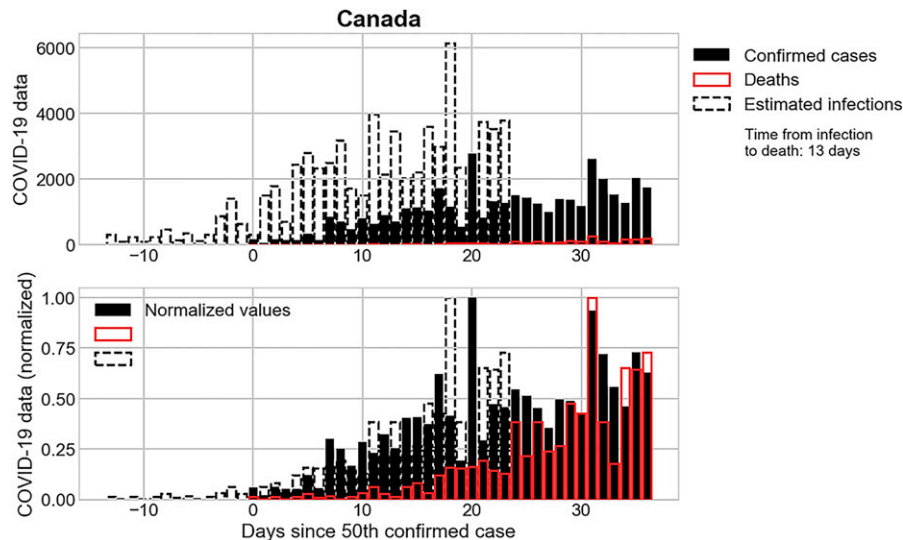


Fig. 1-B

Figs. 1-A and 1-B Charts illustrating successive waves of infections (dashed), detections (black), and deaths (red) for both the U.S. (**Fig. 1-A**) and Canada (**Fig. 1-B**); the x axis is in days. COVID-19 data are also normalized for better visualization. A visual inspection indicates the temporal delay between the waves.

including the U.S. (13,067), U.K. (8,248), Italy (25,028), France (7,103), Spain (19,896), Canada (16,220), and India (335)². However, the extent of testing is not just a policy matter but also is dependent on the availability of scarce public and private resources. Under such circumstances, it may not be prudent for policymakers to rely only on “observable” data (i.e., confirmed COVID-19 cases) as such measures are likely to under-report the extent of the problem. For example, by publicly reporting 47,676 deaths against only 840,476 cases, the United States may not be accounting for the influence of lower levels of testing (13,067 tests per million) relative to other countries. By not proactively acknowledging data that are unobservable—i.e., expected infections that have not been captured by WHO-established COVID-19-detection protocols—policymakers could

be grossly underestimating the true number of infections in the population. Furthermore, if case fatality rates (that is, the ratio of deaths to reported cases; e.g., ~5.7% for the U.S.) do not factor in unobservable infections, models may overestimate the risk of death⁷.

Given this background, we modeled unobserved infections to examine the extent to which we might be grossly underestimating COVID-19 infections in North America.

Materials and Methods

We developed a machine-learning model to uncover hidden patterns based on reported cases and to predict potential infections. First, our model relied on dimensionality reduction to identify parameters that were key to uncovering

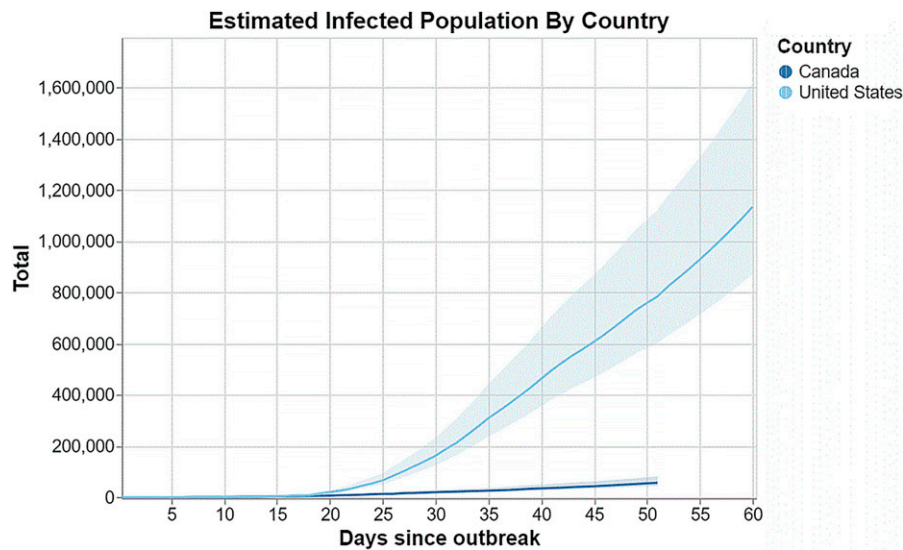


Fig. 2-A

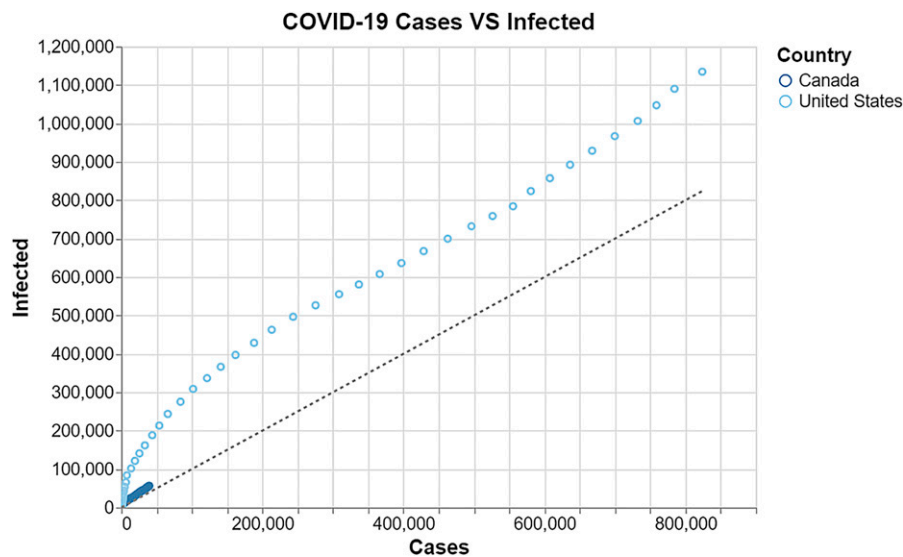


Fig. 2-B

Figs. 2-A and 2-B Line graphs comparing estimated trends of COVID-19 infections in the U.S. and Canada. **Fig. 2-A** The upper and lower bounds are based on hierarchical Bayesian simulations of the case fatality rate. **Fig. 2-B** Relative differences between the U.S. and Canada. As Duarte indicates⁹, deviation from the dashed (45°) line helps compare how the U.S. and Canada have been tracking the true number of infected people.

hidden patterns. Next, our predictive analysis used an unbiased estimator approach to infer past infections from current fatalities.

Open Science

We referenced the initial rapid research and contributions by Pueyo, Duarte, and others⁶⁻¹⁰. Broadly speaking, our analysis compared the numbers of confirmed cases, deaths, and estimated infections across North America (U.S. and Canada). Our data were made available thanks to the generosity of the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), the Esri Living Atlas team, and the Johns Hopkins University Applied Physics Laboratory (JHU APL). The data were

pulled from the COVID-19 Data Repository by the JHU CSSE every hour.

Dimensionality Reduction

We started with exploratory data analysis. We aggregated the U.S. and Canada since they were split by states and provinces. While we focused on North America, we also included countries with a minimum of 10 cases. The columns in our time-series dataset initially included country, state, and number of deaths and confirmed cases. First, we filtered data to include at least 100 cases. As the “deaths” and “confirmed cases” data were separate in the JHU CSSE database, we concatenated both time series. After dropping duplicated columns, we created new

columns for (1) dirty ratio = confirmed cases/(deaths + 1) and (2) fatality = deaths/confirmed cases. We also created new columns for new deaths and new confirmed cases. Next, we created a smaller dataset with at least cumulative 50 cases. We also grouped data for states (U.S.) and provinces (Canada). Our final dataset included the following columns: index, country, date, confirmed cases, deaths, dirty ratio, fatality, new deaths, new confirmed cases, and days since case 50. We assumed that the average time from infection to death could be 8 days (to account for older patients), 13 days, or 23 days (to account for younger patients); however, we based our results on 13 days¹¹. The time from infection to death was taken as being equal to the incubation period plus the time from symptoms to death. This assumption was used to estimate the timing of the infections that led to the observed deaths. We used the fatality rate per country (total deaths/total cases) to estimate the number of infections that were responsible for the observed deaths.

Predictive Analysis

Next, we extrapolated available information with use of Duarte's elegant unbiased hierarchical Bayesian estimator approach^{9,10}. We inferred past infections from current daily deaths as the average fatality rate was approximated from confirmed cases^{9,10,12}. In doing so, we assumed that (1) the case fatality rate (that is, the fatality rate among patients with confirmed cases) may be a good proxy for the fatality rate of the infected population, (2) the growth rate of the infected population may serve as an unbiased estimate of confirmed cases, and (3) on average, the interval between the initial symptoms and death is 8 to 23 days.

Results

Our analysis indicated that, with a 13-day lag time from infection to death, the U.S. likely had at least 1.3 million undetected infections as of April 22, 2020 (represented by dashed bars in Figs. 1-A and 1-B; COVID-19 data are also normalized for better visualization). Under a longer lag time of 23 days, there could have been at least 1.7 million undetected infections. Given these same assumptions, the number of undetected infections in Canada could have ranged from 60,000 to 80,000. We used Duarte's elegant unbiased hierarchical Bayesian estimator approach (which uses the case fatality rate to infer the fatality rate of the unobservable population) as a robustness check on these results. Using that approach, we found that, as of April 22, 2020, the United States had up to >1.6 million undetected infections and Canada had at least 60,000 to 86,000 undetected infections (Figs. 2-A and 2-B). However, the JHU CSSE data feed on April 22, 2020, reported only 840,476 and 41,650 confirmed cases for the United States and Canada, respectively.

Discussion

Our research explored the role of unobservable infections in COVID-19 detection bias. We identified 2 key findings: (1) as of April 22, 2020, the United States may have had 1.5 to 2.02 times the number of reported infections and Canada may

have had 1.44 to 2.06 times the number of reported infections and (2) even if we assume that the fatality and growth rates in the unobservable population (undetected infections) are similar to those in the observable population (confirmed infections), the number of undetected infections may be within ranges similar to those described above. In summary, 2 different approaches indicated similar ranges of undetected infections in North America.

Our analysis has unique strengths. Our multidimensional analysis of unobservable infections in the context of the COVID-19 pandemic has helped us to uncover trends that are likely to have an impact on scientific research in 3 respects. First, we estimated the distribution of successive waves of infections (dashed bars), detections (black bars), and deaths (red bars) for both the U.S. and Canada (Figs. 1-A and 1-B). With the time from infection to death being defined as the incubation period plus the time from symptoms to death, we estimated the timing of the infections that led to the observed deaths. Our results indicated that, as of April 22, 2020, the U.S. may have had at least 1.3 million undetected infections and Canada may have had at least 60,000 undetected infections. Next, we supported these results through a robustness check that used the case fatality rate to infer the fatality rate of the infected population. Assuming that the growth rate of the infected population could be an unbiased estimate of confirmed cases, we found that the numbers of undetected infections may be quite high in the U.S. (>1.6 million) and Canada (60,000 to 86,000) (Figs. 2-A and 2-B).

We also extended the mandate of this research to understand why a set of Western countries accounted for a large number of fatalities despite high testing. As of April 9, 2020, the following countries had relatively higher testing per million population: U.S. (13,067), U.K. (8,248), Italy (25,028), France (7,103), Spain (19,896), and Canada (16,220). Yet, together they accounted for close to 70% of all fatalities. One reason for this finding could be that testing was late as it followed cumulative deaths and new deaths.

Nevertheless, this research is not without shortcomings. To some extent, our method is limited in that asymptomatic carriers (who are believed to account for 30% to 50% of all cases^{13,14}) cannot be observed without antibody tests and thus are not factored into the derived fatality rate. This means that the number of actual infections (dashed bars) is limited to the estimation from observed deaths and cases only and serves as a lower boundary estimate¹⁵ (Figs. 1-A and 1-B). Furthermore, we relied on the extant literature, some of which may still not be peer-reviewed, to arrive at a novel framework that may point to the extent of unobservable infections across North America, specifically, the U.S. and Canada.

However, we hope that readers will appreciate the rapid rate at which the pandemic scenario has evolved over the past weeks and understand the limitations of this research while also acknowledging that unusual times call for unusual solutions. Our goal is to contribute to the ongoing debate on detection bias and to present an alternative mechanism that can help to improve the robustness of COVID-19 data being made available to the scientific community. In summary, our research adds another perspective to the ongoing debate on the

pandemic. However, we highlight the need for more robust data. As the COVID-19 pandemic progresses, it is crucial for policymakers to begin to focus on the potential for detection bias. We must be aware of the extent to which unobservable data—infections that have still not been captured by the system—can damage efforts to “flatten” the pandemic’s curve. ■

Shashank Vaid, PhD¹
Caglar Cakan, PhD(Candidate)²
Mohit Bhandari, MD, PhD³

¹DeGroote School of Business, McMaster University, Hamilton, Ontario, Canada

²Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany

³Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, Ontario, Canada

Email address for M. Bhandari: bhandam@mcmaster.ca

ORCID iD for S. Vaid: [0000-0003-0658-2016](https://orcid.org/0000-0003-0658-2016)

ORCID iD for C. Cakan: [0000-0002-1902-5393](https://orcid.org/0000-0002-1902-5393)

ORCID iD for M. Bhandari: [0000-0003-3556-9179](https://orcid.org/0000-0003-3556-9179)

References

1. Chan SP. Coronavirus: 'World faces worst recession since Great Depression.' BBC. 2020. Accessed 2020 Apr 21. <https://www.bbc.com/news/business-52273988>
2. Worldometer. Data on COVID-19 coronavirus pandemic (2020). 2020. Accessed 2020 Apr 21. <https://www.worldometers.info/coronavirus/>
3. Normille D. 'Suppress and lift': Hong Kong and Singapore say they have a coronavirus strategy that works. Science Magazine. 2020 Apr 13. Accessed 2020 Apr 21. <https://www.sciencemag.org/news/2020/04/suppress-and-lift-hong-kong-and-singapore-say-they-have-coronavirus-strategy-works>
4. Hale T, Webster S, Petherick A, Phillips T, Kira B. Oxford COVID-19 government response tracker. Blavatnik School of Government, University of Oxford. 2020. Accessed 2020 Apr 21. <https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker>
5. World Health Organization. National capacities review tool for a novel coronavirus. 2020 Jan 9. Accessed 2020 Apr 21. <https://www.who.int/publications-detail/national-capacities-review-tool-for-a-novel-coronavirus>
6. Lipsitch M. Estimating case fatality rates of COVID-19 Lancet Infect Dis. 2020 Mar 31. [Epub ahead of print].
7. Xu Z, Li S, Tian S, Li H, Kong LQ. Full spectrum of COVID-19 severity still being depicted. Lancet. 2020 Mar 21;395(10228):947-8. Epub 2020 Feb 14.
8. Pueyo T. Coronavirus: why you must act now. Medium. 2020 Mar 10. Accessed 2020 Apr 21. <https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca>
9. Duarte JB. Codes. Accessed 2020 Apr 21. <http://jbduarte.com/codes/>
10. Flaxman S, Mishra S, Gandy A, Unwin HJT, Coupland H, Mellan TA, Zhu H, Berah T, Eaton JW, Guzman PNP, Schmit N, Cilloni L, Ainslie KEC, Baguelin M, Blake I, Boonyasiri A, Boyd O, Cattarino L, Ciavarella C, Cooper L, Cucunubá Z, Cuomo-Dannenburg G, Dighe A, Djaafara B, Dorigatti I, van Elsland S, FitzJohn R, Fu H, Gaythorpe K, Geidelberg L, Grassly N, Green W, Hallett T, Hamlet A, Hinsley W, Jeffrey B, Jorgensen D, Knock E, Laydon D, Nedjati-Gilani G, Nouvellet P, Parag K, Siveroni I, Thompson H, Verity R, Volz E, Walters C, Wang H, Wang Y, Watson O, Winskill P, ZI Z, Whittaker C, Walker PGT, Ghani A, Donnelly CA, Riley S, Okell LC, Vollmer MAC, Ferguson NM, Bhatt S. Report 13: estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. 2020 Mar 30. Accessed 2020 April 21. <https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Europe-estimates-and-NPI-impact-30-03-2020.pdf>
11. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Pastore Y, Piontti A, Mu K, Rossi L, Sun K, Viboud C, Xiong X, Yu H, Halloran ME, Longini IM, Jr, Vespignani A. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak Science. 2020 Mar 6. [Epub ahead of print].
12. Feenstra RC, Inklaar R, Timmer MP. The next generation of the Penn World Table. Am Econ Rev. 2015;105(10):3150-82.
13. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet. 2020 Apr 11; 395(10231):1225-8. Epub 2020 Mar 13.
14. Freund A. Up to 30% of coronavirus cases asymptomatic. DW News. 2020 Mar 24. Accessed 2020 Apr 21. <https://www.dw.com/en/up-to-30-of-coronavirus-cases-asymptomatic/a-52900988>
15. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Euro Surveill. 2020 Mar;25(10): 2000180.